DOCUMENT RESUME

ED 302 573 TM 012 585

AUTHOR Ferrara, Steven F.
TITLE Effects of Essay Order on Raters' Score Assignments in a Large-Scale Writing Assessment.
PUB DATE 18 Apr 87
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Essay Tests; Evaluators; *Graduation Requirements; High Schools; High School Students; *Interrater Reliability; *Scoring; Standardized Tests; State Programs; Test Bias; Testing Programs; *Writing Evaluation
IDENTIFIERS *Context Effect; *Maryland Functional Writing Test; Rater Reliability; Test Order

ABSTRACT

The necessity of controlling the order in which trained essay raters for a statewide writing assessment program receive student essays was studied. The underlying theoretical question concerns possible rater bias caused by raters reading long strings of essays of homogeneous quality; this problem is usually referred to as context effect or contrast effect. The program involves the Maryland Writing Test (MWT), which is administered each year to high school students (grades 9 through 12) as a graduation requirement. Each essay in the test is scored by two raters on a four-point scale using a holistic procedure. A sample of 1,000 MWT essay booklets from over 85,000 student booklets was selected. Raters (23 narrative and 15 explanatory) who participated in the live scoring of the 1986 MWT test program scored 957 pairs of essays. Three ordering schemes were assessed. There were 433 double readings of essays on one topic, and 652 double readings of essays on another topic. Analysis of variance on resulting data indicated that the order in which essays were placed into scoring packets did affect the raters' score assignments, at least for the narrative essay prompt used in the study. The same conclusion cannot be supported for the explanatory prompt. Other results indicate that rater agreement rates and pass-rates do not differ by essay sequence, and that some raters are more reliable scorers than are others within and across sequence conditions. Five data tables are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED302573

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

STEVEN FERRARA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Effects of Essay Order on Raters' Score Assignments in a Large-Scale Writing Assessment

Steven Ferrara
Maryland State Department of Education

Paper presented at the Annual Meeting of the American Educational Research Association, April 21, 1987, Washington, DC

TM012 585

The purpose of this study is to determine if it is necessary to control the order in which trained essay raters for a statewide writing assessment program receive student essays. The underlying theoretical question concerns possible rater bias caused by raters reading long strings of essays of homogeneous quality, usually referred to as context effect (Hughes & Keeling, 1984; Hughes, Keeling, & Tuck, 1980a, 1980b, 1983) and contrast effect (Daly & Dickson-Markman, 1982). However, the practical question for a statewide writing assessment program, like that of the Maryland State Department of Education, is whether it is necessary to shuffle the sequential order of essays as they are collected from classrooms and schools and assembled into packets prior to scoring.

Background

Several states now administer direct writing assessments as a part of their statewide competency-testing programs. One such example is the Maryland Writing Test (MWT), one of four basic skills tests in the Maryland Functional Testing Program used to diagnose students' strengths and weaknesses and to certify them for high school graduation. Large-scale writing programs face various practical problems that require expensive and time-consuming solutions. One such problem stems from the logistics of test distribution and retrieval: each school and school district receives a long series of consecutively numbered essay test booklets from the entire collection of consecutively numbered booklets. For example, District A might receive booklets numbered 1 to 1,000, District B might receive booklets numbered 1,001 to 2,000, and so on. A potential problem arises when this sequential order is preserved and essay booklets are distributed to raters: under this condition individual raters would receive a long series of booklets from the same school and district. A potential for biasing raters arises because essays from the same school tend to display similar writing strategies and to be of somewhat similar low, medium, or high quality. These similarities are due in part to the uniformity of writing instruction, general achievement levels, and student populations within schools and districts.

Several large-scale writing assessment experts around the country that we canvassed believe that typical systems for packaging essays (into scoring packets of 25 essays) and distributing packets to raters sufficiently breaks up sequential ordering of essays for individual raters to control for contrast effect bias (cf, Daly & Dickson-Markman, 1982; White, 1985, pp. 158-159).

However, writing assessment researchers have shown that raters' perceptions of the quality of individual essays are affected by the context in which they read an essay (Daly & Dickson-Markman, 1982; Hales & Tokar, 1975; Hughes & Keeling, 1984; Hughes et al., 1980a, 1980b, 1983). That is, it has been shown that raters may be biased upward after scoring a large number of poorly written essays or biased downward after reading a large number of well written essays. For example, it is possible that a rater may become inclined to rate a "good" essay as "excellent" after reading a long string of "poor" essays because of the contrast of quality between the good essay and preceding string of poor essays.

To control for contrast effect bias, the sequential order of essays is often shuffled randomly. The costs to shuffle MWT booklets (which contain two essays per student) can be quite high: thousands of dollars to shuffle over 80,000 essay booklets and a week's delay before scoring can begin, and thousands more dollars to return booklets to numerical order and a month or more delay before essay booklets can be returned to students. These costs are not easily justifiable, given time and fiscal constraints of statewide writing assessment programs, unless random ordering of essays is demonstrated to be necessary to control contrast effect bias. In addition, demonstration that random ordering is necessary to avoid this bias is important information for large-scale writing assessment programs that are not currently attending to this concern.

This study was designed and conducted to investigate the question Are essay scores assigned by trained professional raters influenced by sequence of essays? Three factors that may affect scores assigned to essays by raters are tested: essay sequence (random, "mixed-up," and sequential), school district from which essays are sampled, and individual rater. Our hypothesis is that there is no difference in scores assigned to essays due to sequence. However, we do expect differences due to school district since the general and writing achievement levels of districts in any state vary.

Method

Instrument

The MWT is administered once each school year to ninth graders and above. Students write a narrative essay on the first day and an explanatory essay on the second, both in response to prompts that have been developed, field tested, and selected for use in the MWT. For the MWT, each essay is scored by two raters on a four-point scale using the Maryland Department of Education's modified holistic scoring

procedure. The two ratings on an essay must be in either exact or adjacent agreement or resolved by a third, expert reader. An essay's holistic score is the average of two ratings. A student's total holistic score on the MWT is the sum of these two average scores. The maximum possible total holistic score, thus, is 8.0; the passing total holistic score on the MWT is 5.5. Pass-rates on the MWT (that is, the percentage of ninth graders who pass the test on first try) have climbed steadily from 51.4 percent in 1984 to 68.5 percent in 1986.

Subjects

Student essays. We selected a sample of 1,000 MWT essay booklets from over 85,000 student booklets. We constructed the sample by first randomly sampling districts to represent important regions in the state and to approximate the proportions of students in the entire state in urban, suburban, and rural districts. Finally, schools within districts were selected to meet these initial selection criteria and to approximate the statewide pass-rate on the 1986 MWT.

Raters. Raters (23 Narrative, 15 Explanatory) who participated in the live scoring of the 1986 MWT scored 957 pairs of essays for the study. They were retrained in the use of the MWT modified holistic scoring rubric because regular scoring had been completed three months prior to the present study. All raters scored some essays within each of the three sequence conditions. However, because the sequencing of essays is different in each condition, raters encountered the same essay in more than one sequence only a small percentage of times. In addition, it was not possible for raters to encounter the same essay within sequence condition because packet order is counter-balanced (see Table 1 below). The design for crossing rater and sequence is contained in Table 1.

Table 1

Design for Crossing Raters with Essay Sequence (Within Essay Topic)

	Packet Numbers/Reading		
Raters	Numerical	Mixed-Up	Random
Group 1	1-14/ First	15-27/ First	1-14/ Second
	15-27/ Second	28-40/ Second	28-40/ First
Group 2	1-14/ Second	1-14/ First	15-27/ First
	28-40/ First	15-27/ Second	28-40/ Second
Group 3	15-27/ First	1-14/ Second	1-14/ First
	28-40/ Second	28-40/ First	15-27/ Second

Procedure

The 1,000 pairs of essays were duplicated, split into narrative and explanatory stacks, and placed into packets in three different sequences:

1. Numerical order. Essays were placed in packets in their original numerical order. In addition, packets remained in their numerical order as they were distributed to raters. As a result, raters read strings of essays within packets from the same district.

2. Mixed-up order. Essays in this condition were placed in five stacks (one per district). The top 1/40th of the essays in each stack were placed into a packet, one stack at a time, in order of stacks. In this condition, raters encountered consecutive strings of essays from the same district but essays from all five districts within each packet. Packets were distributed to raters in a non-systematic mixed-up order.

3. Random order. In this condition, the numerical order of essays was shuffled randomly prior to placing essays in packets. Packet constructors used slips of paper, each with 25 district numbers, to order essays within packets. District numbers were sampled randomly and in proportion to the percentage of essays from each district included in the study. As a result, raters saw only occasional consecutive strings of essays from the same school district in a packet and saw essays from all five districts in each packet.

Six sets of scoring packets -- three sets of narrative packets with duplicated essays placed into the packets in the three different sequences; likewise, three sets of explanatory essays -- were then distributed to six "groups" of raters for scoring (see Table 1). All ratings were assigned independently (except, possibly, when raters encountered the same essay for a second time in a different sequence; see below). The six scores for each essay were then key-entered into data files and the 12 scores assigned to each student's pair of narrative and explanatory essays were merged into a single data file for subsequent analysis.

Results and Discussion

We conducted three sets of analyses to examine the effects of essay sequence on raters' score assignments. The first focuses on rater agreement rates within and across the three essay sequences. This analysis includes a comparison of percentages of exact and adjacent agreement and disagreement. The second set of analyses examines effects of sequence, district, and reader on individual ratings via ANOVA. The final analysis compares student pass-rates across the three sequences.

Analyses of these data indicate no significant differences in individual ratings for each essay, rater agreement rates, or pass-rates due to essay sequence. However, ratings did differ by the district from which essays were selected, as expected, and both by rater and by rater within sequence in essay topic one. Analyses and results are described in detail below.

Rater Agreement

We first examined rater agreement rates within and across sequences for each topic separately. Table 2 contains the percents of times that two ratings for an essay were exactly equal both within and across sequences. As shown in Table 2, agreement rates within sequences are virtually identical with rates across sequences for both narrative and explanatory essays. In addition, these rates

are consistent with exact agreement rates obtained in live scoring (see Table 2 below).

Table 2

Percent of Exact Rater Agreement Within and Across Sequence for Topics One and Two

Sequence	1	2	3
Topic 1			
1. Numerical	77	77	75
2. Mixed-up		74	73
3. Random			72
Topic 2			
1. Numerical	70	68	68
2. Mixed-up		69	68
3. Random			68

Note. Percentages on diagonals are based on 957 pairs of readings of individual essays within sequence; percentages off diagonals are based on four sets of 957 pairs of readings of essays across sequences.

As is typical for many large-scale, direct writing assessment programs, ratings on a single essay must be in exact or adjacent (one point difference) agreement for the MWT. An essay that receives two ratings that differ by more than one point receives a third rating by an expert reader. This third rating is averaged with the score from the first two ratings with which it is closest. Of the 6,067 pairs of readings (on each of two essays for students in three sequences), only 39 (0.6 percent) required a third reading resolution. The frequencies of third readings are spread evenly across sequences but unevenly across topics: eight third reader resolutions were required for narrative essays, 31 for explanatory essays. Rater disagreement appears not to be a significant factor in this study.

Our design crosses all raters with all sequences so that rater by sequence interactions can be examined. As a result of this crossing, some raters encountered and read the same essay twice. However, this double reading could occur only across sequence but not within sequence (see Table 1 above). Of the 5,742 total readings (2 readings for

957 paired essays across 3 sequences), there were 433 (7.5%) double readings of topic one essays and 652 (11.4%) double readings of topic two essays. To look for effects of these double readings and to assess the validity of ratings of essays read twice by the same reader, we compared rates of exact and adjacent agreement and non-agreement within and across sequence. Agreement and non-agreement rates are contained in Table 3 below. Agreement and non-agreement rates within sequence are virtually the same as those across sequence (see Table 3), indicating that the double readings that did occur due to crossing raters with sequences did not significantly impact score assignments.

Table 3

Percent of Agreement and Non-Agreement Within and Across Sequence* for Topics One and Two

	Exact	Adjacent	Non-Agreement
	Topic 1		
Within	74.3	25.5	0.2
Across	74.4	25.4	0.2
	Topic 2		
Within	69.0	30.0	1.0
Across	67.9	31.3	0.8

*Double readings possible only across sequence (see Table 1 for rater by sequence design).

Note. Agreement rates within sequence are based on 3 sets of 957 pairs of readings; rates across sequence are based on 12 sets of 957 pairs of readings.

Effects on Individual Ratings

Tables 4 and 5 contain results of ANOVAs within essay topic. The design for the results in Tables 4 and 5 is a three-way mixed effects ANOVA with two ratings on each essay as the dependent variable. Because our goal is to generalize results to all trained raters and all districts in Maryland, rater and district are treated as random factors; sequence is treated as a fixed factor. Sums of squares in the tables are calculated using SAS PROC GLM;

mean squares are calculated from the Type III (or partial) sums of squares for unbalanced designs; the F ratios are hand-calculated ratios for mixed effects designs using guidelines in Ostle (1963, p. 327).

The most important results in Tables 4 and 5 for the purposes of the present study are the non-significant F-tests for the sequence effect but the significant rater and sequence by rater (topic one only) effects (see Tables 4 and 5). These results suggest that (1) Some raters are more reliable scorers than others both within and across sequence conditions; and (2) Some raters' score assignments may be affected by the sequence of essays. Regarding reliability differences among raters, all raters who participate in scoring of MWT essays, both in live scoring and in this study, must qualify to participate in scoring by reaching 80 percent agreement or higher with pre-scored essays in qualifying rounds. In addition, the raters in this study had participated in live scoring of MWT essays in 1986 and some in previous years. Despite the quality of raters in the study, however, trained professional raters do differ in reliability. One indicator of the degree of unreliability is the percentage of essays on which raters' scores differ by more than one point (referred to as non-agreement). The range of disagreement rates for all raters in the present study is 0.2 to 1.3 percent, well within the bounds of acceptability and similar to ranges for live scoring. This difference between the most and least reliable raters, as small as it is, probably accounts in part for the effects on explanatory ratings due to rater. Also, adjacent agreement -- that is, accepted one-point differences between raters -- probably accounts in part for rater effects. However, interpretation is not so simple for narrative ratings because of the significant sequence by rater interaction (see Table 4). This interaction indicates that raters' score assignments differ as a function of essay sequence. It also suggests that rater reliability may be affected by essay sequence, at least for narrative essays.

Table 4

ANOVA for Effects of Sequence, Rater, and District on Individual Essay Ratings in Topic One

Source	df	MS	F	p
Sequence (S)	2	0.84	2.28	.100
Rater (R)	22	0.86	2.36	.005
District (D)	4	74.45	203.98	.000
S x R	25	0.70	1.90	.025
S x D	8	0.27	0.73	.250
R x D	82	0.37	0.81	.250
S x R x D	66	0.37	0.81	.250
Error	5,783	0.45		

Table 5

ANOVA for Effects of Sequence, Rater, and District on Individual Essay Ratings in Topic Two

Source	df	MS	F	p
Sequence (S)	2	0.06	0.15	.250
Rater (R)	14	1.70	2.82	.010
District (D)	4	90.81	150.60	.000
S x R	28	0.63	1.47	.100
S x D	8	0.09	0.20	.250
R x D	56	0.60	0.98	.250
S x R x D	81	0.43	0.69	.250
Error	5,876	0.62		

One additional point should be made about the results in Tables 4 and 5. Effects on ratings due to district are significant for both topic one and topic two essays (see Tables 4 and 5). This result indicates that there are significant differences in the ratings -- and the quality of writing -- of essays from different districts, as expected.

Effects on Pass-Rates

The effect of essay sequence on student pass-rates is of ultimate importance for the MWT because the test is used to grant and deny high school diplomas. The question addressed in this analysis regards whether students would be likely to pass the MWT if essays were distributed to raters in one sequence but likely to fail if essays were distributed to raters using a different sequence.

The pass-rates for each sequence are 56.7 percent for numerical, 54.9 percent for mixed-up, and 57.5 percent for random. The differences between these pass-rates are not significant. The z-values for the numerical versus mixed-up sequences is .793 ($p \leq .206$); for numerical versus random is .354 ($p \leq .337$); and for mixed-up versus random, 1.15 ($p \leq .111$). Thus, there appears to be no significant effect on pass-rate due to the order in which raters receive essays.

Conclusions

Results of the ANOVAs on essay scores indicate that the order in which essays are placed into scoring packets -- and, thus, the order in which raters receive essays -- does affect raters' score assignments, at least for the narrative essay prompt used in this study. The same conclusion cannot be drawn for the explanatory prompt used in this study. In addition, other results reported here indicate that rater agreement rates and pass-rates do not differ by essay sequence.

Earlier it was argued that rater effects may be due to small but influential differences among raters in reliability. It was also suggested that the range of non-agreement is well within typical and acceptable limits. In terms of score reliability, then, the significant rater effects are not of great concern. In terms of interpreting this study and making decisions about sequencing essays for future live scoring, however, rater and sequence by rater effects are cause for some consideration. Since rater agreement rates are adequate, the rater by sequence effect may be considered as a concern for score validity. That is, contrast effect bias appears to be operating for some raters in some essay sequence conditions. This finding is partly consistent with earlier studies of contrast effect which did not use trained raters and which did not simulate large-scale writing assessment scoring conditions and essay sequences (see Daly & Dickson-Markman, 1982).

The non-significant results for the effects due to sequence by rater for explanatory essays may reflect the ability of trained professional raters to disregard extraneous factors during scoring such as contrast effect. This is similar to the ability required when court room juries are instructed to disregard specific remarks, evidence, or viewpoints. A further implication of these results is that rater training in the use of a scoring guide can offset the influence of contrast effect. This conclusion runs contrary to findings from previous studies using untrained, "amateur" raters, such as teachers, whose score assignments were influenced by contrast effect bias (Daly & Dickson-Markman, 1982). It is also inconsistent with other studies that suggest that, in general, essay

raters tend to be influenced by irrelevant features of essays such as penmanship (Chase, 1986) and length of essay and other factors (Charney, 1984).

The generalizability of these results to (1) all essays written to future MWT narrative and explanatory prompts, (2) other large-scale writing assessment programs, and (3) other standardized direct writing assessments is limited by several features of the present study's design. These features include the degree to which: the five schools/districts sampled for the study represent relevant features for the entire state and beyond; the procedures for sequencing essays from five schools/districts for the study replicate the procedures for essays from 24 school districts; and training and scoring procedures and conditions replicate those for live scoring. Other factors that may limit the generalizability of these results are related to the unique features of the MWT: its two-essay format, Maryland's version of modified holistic scoring and training, and the nuances of its scoring rubrics.

Further investigation into the issue of contrast effects in scoring essays from large-scale writing assessments should include effects on other modes of discourse and several prompts within each essay to determine if contrast effects generalize across modes and prompts or vary according to modes and prompts. Follow-up to this study might include determination of which raters and essay sequences interacted; some evidence, perhaps anecdotal, of the causes of this interaction; and investigation of the lack of interaction for explanatory essays.

Finally, the results of this study suggest that it is probably not reasonable to forego expensive and time-consuming procedures for randomizing essays prior to scoring packet construction until further and clearer results can be obtained. However, if time and fiscal constraints are prohibitive, a prudent and relatively inexpensive alternative may be to use some version of the "mixed-up" sequence described in this study.

References

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. Research on the Teaching of English, 18, 65-81.

Chase, C. I. (1986). Essay test scoring: Interaction of relevant variables. Journal of Educational Measurement, 23, 33-41.

Daly, J. A., and Dickson-Markman, F. (1982). Contrast effects in evaluating essays. Journal of Educational Measurement, 19 (4), 309-316.

Hales, L. W., & Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. Journal of Educational Measurement, 12, 115-117.

Hughes, D. C., & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. Journal of Educational Measurement, 21 (3), 277-281.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1980a). Essay marking and the context problem. Educational Research, 22, 147-148.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1980b). The influence of context position and scoring method on essay scoring. Journal of Educational Measurement, 17, 131-135.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1983). The effects of instructions to scorers intended to reduce context effects in essay scoring. Educational and Psychological Measurement, 43, 1047-1049.

Ostle, B. (1963). Statistics in research (2nd ed.). Ames, Iowa: Iowa State University Press.

White, E. M. (1985). Teaching and assessing writing. San Francisco, Jossey-Bass.

b:sf2001:4/18/87